日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。
This is to certify that the annexed is a true copy of the following application as filed th this Office.

出願年月日
Date of Application: ２００３年　７月１７日

出願番号
Application Number: 特願２００３－２７５９８３
ST. 10/C]: [ＪＰ２００３－２７５９８３]

願人
Applicant(s): 日本電気株式会社
大瀧　慈
社団法人バイオ産業情報化コンソーシアム

２００３年１０月２９日

特許庁長官
Commissioner,
Japan Patent Office 今井康夫

出証番号　出証特２００３－３０８９５５

PATENT APPLICATION

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of

Masataka ANDOH, et al.

Docket No: Q79665

Appln. No.: 10/766,011

Group Art Unit: 1631

Confirmation No.: 2008

Examiner: Unknown

Filed: January 29, 2004

For: SYSTEM, METHOD, AND PROGRAM FOR ESTIMATING GENE EXPRESSION STATE, AND RECORDING MEDIUM THEREFOR

SUBMISSION OF PRIORITY DOCUMENT

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

Submitted herewith is a certified copy of the priority document on which a claim to priority was made under 35 U.S.C. § 119. The Examiner is respectfully requested to acknowledge receipt of said priority document.

Respectfully submitted,

SUGHRUE MION, PLLC
Telephone: (202) 293-7060
Facsimile: (202) 293-7860

WASHINGTON OFFICE
23373
CUSTOMER NUMBER

Howard L. Bernstein
Registration No. 25,665

Enclosures: Japan 2003-275983

Date: August 12, 2004

【書類名】　特許願
【整理番号】　64003001
【特記事項】　特許法第３０条第１項の規定の適用を受けようとする特許出願
【提出日】　平成15年 7月17日
【あて先】　特許庁長官殿
【国際特許分類】　G06N　7/00
【発明者】
　【住所又は居所】　東京都港区芝五丁目７番１号　日本電気株式会社内
　【氏名】　安東　正貴
【発明者】
　【住所又は居所】　東京都港区芝五丁目７番１号　日本電気株式会社内
　【氏名】　斎藤　彰
【発明者】
　【住所又は居所】　広島県廿日市市宮園９丁目１の７
　【氏名】　大瀧　慈
【発明者】
　【住所又は居所】　広島県広島市佐伯区楽々園５－９　五日市住宅１６－２０２
　【氏名】　佐藤　健一
【発明者】
　【住所又は居所】　広島県広島市南区仁保南２丁目８－７
　【氏名】　西山　正彦
【発明者】
　【住所又は居所】　東京都中央区八丁堀二丁目２６番９号　グランデビルディング
社団法人バイオ産業情報化コンソーシアム内
　【氏名】　大谷　敬子
【特許出願人】
　【識別番号】　000004237
　【氏名又は名称】　日本電気株式会社
【特許出願人】
　【識別番号】　503077165
　【氏名又は名称】　大瀧　慈
【特許出願人】
　【識別番号】　500535301
　【氏名又は名称】　社団法人バイオ産業情報化コンソーシアム
【代理人】
　【識別番号】　100071272
　【弁理士】
　【氏名又は名称】　後藤　洋介
【選任した代理人】
　【識別番号】　100077838
　【弁理士】
　【氏名又は名称】　池田　憲保
【手数料の表示】
　【予納台帳番号】　012416
　【納付金額】　21,000円
【その他】　国等の委託研究の成果に係る特許出願（平成１４年度新エネルギー・産業技術総合開発機構「遺伝子多様性モデル解析事業」委託研究、産業活力再生特別措置法第３０条の適用を受けるもの）
【提出物件の目録】
　【物件名】　特許請求の範囲　1

【物件名】　明細書　1
【物件名】　図面　1
【物件名】　要約書　1
【包括委任状番号】　0018587
【包括委任状番号】　0307208

【書類名】特許請求の範囲
【請求項１】
遺伝子発現強度データを送出する入力装置と、プログラム制御により動作するデータ解析装置と、出力装置とを含み、チャンネル毎に遺伝子が発現している確率を推定する遺伝子発現状況推定システムにおいて、
前記データ解析装置は、
前記入力装置から送出された遺伝子発現強度データを用いて、以下の数１に示される混合正規分布（尚、外１は平均０，分散σ^2の１次元正規分布の密度関数を表し、外２－１及び外２－２はそれぞれ第１および第２のコンポーネントの平均と分散パラメータを表し，ξは混合率を表し、外３が満たされているものとする。）の分布パラメータを推定し、この推定された分布パラメータを送出する分布パラメータ推定手段と、
【数１】

$$(1-\xi)\phi\left(u-\mu_0 \mid \sigma_0^2\right)+\xi\phi\left(u-\mu_1 \mid \sigma_1^2\right)$$

[外１]

$$\phi\left(* \mid \sigma^2\right)$$

[外２－１]

$$\left(\mu_0,\ \sigma_0^2\right)$$

[外２－２]

$$\left(\mu_1,\ \sigma_1^2\right)$$

[外３]

$$\mu_0 < \mu_1,\ \sigma_0^2 > 0,\ \sigma_1^2 > 0,\ 0 < \xi < 1$$

前記入力装置から送出された遺伝子発現強度データと、前記分布パラメータ推定手段から送出された分布パラメータとを用いて、混合正規分布の混合比パラメータを推定し、この推定された混合比パラメータを送出する混合比パラメータ推定手段と、
前記遺伝子発現強度データと、前記推定された分布パラメータ及び混合比パラメータを用いて、各チャンネル毎に各遺伝子の発現状況に関する事後確率を計算し、計算された事後確率を送出する事後確率計算手段を有し、
計算された前記事後確率が前記出力装置に出力されることを特徴とする遺伝子発現状況推定システム。
【請求項２】
前記分布パラメータ推定手段は、２つのチャンネルの遺伝子発現強度Ｘ，Ｙの差を示すＶが０近傍である遺伝子に関する発現量の和のデータに対して、２つのコンポーネントからなる前記混合正規分布を当てはめ、混合率（ξ）、平均（μ_0，μ_1）、及び分散外３－１を推定する
[外３－１]

$$\left(\sigma_0^2,\ \sigma_1^2\right)$$

ことを特徴とする請求項１記載の遺伝子発現状況推定システム。

【請求項３】

前記遺伝子に関する発現量の和のデータは、遺伝子発現強度Ｘ，Ｙの差の絶対値外４の中央値をc_Mとするとき、外５で示される

[外４]

$$|v_i|(i = 1, \cdots, n)$$

[外５]

$$\{u_i | |v_i| \leq c_M, i = 1, \cdots, n\}$$

ことを特徴とする請求項２記載の遺伝子発現状況推定システム。

【請求項４】

前記分布パラメータ推定手段は、推定された外６を用いて、外７を以下の数２、数３（尚、N_0は外８を満たすデータのインデクッス集合とし、外９はその要素の個数とする。）、数４、及び数５に従って推定する

【数２】

$$\hat{\mu} = (\hat{\mu}_1 - \hat{\mu}_0)/2$$

【数３】

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{2\|N_0\|}\sum_{i \in N_0} v_i^2$$

【数４】

$$\hat{\sigma}_\beta^2 = \frac{1}{4}\hat{\sigma}_0^2 - \frac{1}{2}\hat{\sigma}_\varepsilon^2$$

【数５】

$$\hat{\lambda} = \sqrt{\log\left(1 + \frac{\hat{\sigma}_1^2 - \hat{\sigma}_0^2}{4\hat{\mu}^2}\right)}$$

[外６]

$$\hat{\xi}, \hat{\mu}_0, \hat{\sigma}_0^2, \hat{\mu}_1, \hat{\sigma}_1^2$$

[外７]

$$\mu, \sigma_\varepsilon^2, \sigma_\beta^2, \lambda$$

[外８]

$$i \in \{ i \mid u_i < \hat{\mu}_0 \}$$

[外９]

$$\| N_0 \|$$

ことを特徴とする請求項３記載の遺伝子発現状況推定システム。

【請求項５】

前記混合比パラメータ推定手段は、前記入力装置から送出された遺伝子発現強度データ｛（u_i，v_i）｜ｉ＝１，．．．，ｎ｝に対して、前記分布パラメータ推定手段から与えられた外１０を用いて、以下の数６（ただし、以下の数７及び数８に示される関係を満たすとする。）に示される４つのコンポーネントからなる２変量混合正規分布を当てはめ、混合比パラメータｐ＝（p_{00}，p_{10}，p_{01}，p_{11}）（p_{00}は細胞１，２において、共に遺伝子が発現していない場合における混合比パラメータを表し、p_{11}は細胞１，２において、共に遺伝子が発現している場合における混合比パラメータを表し、p_{10}は細胞１において遺伝子が発現しており、細胞２において遺伝子が発現していない場合における混合比パラメータを表し、p_{01}は細胞１において遺伝子が発現していなく、細胞２において遺伝子が発現している場合における混合比パラメータを表す）を推定する

[外１０]

$$\hat{\theta} = \left(\hat{\mu}, \hat{\lambda}, \hat{\sigma}_\varepsilon^2, \hat{\sigma}_\beta^2 \right)$$

【数６】

$$p_{00} g_{00}(u, v \mid \hat{\theta}) + p_{10} g_{10}(u, v \mid \hat{\theta}) + p_{01} g_{01}(u, v \mid \hat{\theta}) + p_{11} g_{11}(u, v \mid \hat{\theta})$$

$$\cong p_{00} \phi\left(u \mid 4\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\varepsilon^2 \right) \phi\left(v \mid 2\hat{\sigma}_\varepsilon^2 \right) + p_{10} \phi_2\left(u - \hat{\mu}, v - \hat{\mu} \mid \Sigma_{10} \right)$$

$$+ p_{01} \phi_2\left(u - \hat{\mu}, v + \hat{\mu} \mid \Sigma_{01} \right) + p_{11} \phi\left(u - 2\hat{\mu} \mid 4\hat{\mu}^2 (e^{\hat{\lambda}^2} - 1) \right.$$

$$\left. + 4\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\varepsilon^2 \right) \phi\left(v \mid 2\hat{\sigma}_\varepsilon^2 \right)$$

【数７】

$$\hat{\Sigma}_{10} = \begin{pmatrix} \hat{\mu}^2 (e^{\hat{\lambda}^2} - 1) + 4\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\varepsilon^2 & \hat{\mu}^2 (e^{\hat{\lambda}^2} - 1) \\ \hat{\mu}^2 (e^{\hat{\lambda}^2} - 1) & \hat{\mu}^2 (e^{\hat{\lambda}^2} - 1) + 2\hat{\sigma}_\varepsilon^2 \end{pmatrix}$$

【数８】

$$\hat{\Sigma}_{01}=\begin{pmatrix} \hat{\mu}^2(e^{\hat{\lambda}^2}-1)+4\hat{\sigma}_\beta^2+2\hat{\sigma}_\varepsilon^2 & -\hat{\mu}^2(e^{\hat{\lambda}^2}-1) \\ -\hat{\mu}^2(e^{\hat{\lambda}^2}-1) & \hat{\mu}^2(e^{\hat{\lambda}^2}-1)+2\hat{\sigma}_\varepsilon^2 \end{pmatrix}$$

ことを特徴とする請求項４記載の遺伝子発現状況推定システム。

【請求項６】

前記事後確率計算手段は、前記入力装置から送出された遺伝子発現強度データの対（ｕ，ｖ）に対して、細胞１および細胞２における任意の遺伝子が発現している事後確率を、以下の数９（尚、外１０－１は、以下の数１０によって与えられるものとする。）及び数１１（尚、τ_1，τ_2はそれぞれの細胞における遺伝子の真の発現の有無を１および０で表したものである。）にしたがって算出する

【数９】

$$Pr(\tau_1=1\mid\hat{p},\hat{\theta})=\frac{\hat{p}_{10}g_{10}(u,v\mid\hat{\theta})+\hat{p}_{11}g_{11}(u,v\mid\hat{\theta})}{f(u,v\mid\hat{p},\hat{\theta})}$$

【数１０】

$$f(u,v\mid\hat{p},\hat{\theta})=\sum_{(j,k)\in\{0,1\}^2}\hat{p}_{jk}g_{jk}(u,v\mid\hat{\theta})$$

【数１１】

$$Pr(\tau_2=1\mid\hat{p},\hat{\theta})=\frac{\hat{p}_{01}g_{01}(u,v\mid\hat{\theta})+\hat{p}_{11}g_{11}(u,v\mid\hat{\theta})}{f(u,v\mid\hat{p},\hat{\theta})}$$

[外１０－１]

$$f\left(u,v\mid\hat{P},\hat{\theta}\right)$$

ことを特徴とする請求項５記載の遺伝子発現状況推定システム。

【請求項７】

前記事後確率計算手段は、細胞１および細胞２で遺伝子の発現状態が異なる状態である事後確率を、以下の数１２（尚、τ_1，τ_2はそれぞれの細胞における遺伝子の真の発現の有無を１および０で表したものである。）にしたがって算出する

【数１２】

$$Pr(\tau_1\neq\tau_2\mid\hat{p},\hat{\theta})=\frac{\hat{p}_{10}g_{10}(u,v\mid\hat{\theta})+\hat{p}_{01}g_{01}(u,v\mid\hat{\theta})}{f(u,v\mid\hat{p},\hat{\theta})}$$

ことを特徴とする請求項５記載の遺伝子発現状況推定システム。

【請求項８】

前記事後確率計算手段は、全ての遺伝子発現強度データの対（ｕ，ｖ）に基づいて、遺

伝子が発現している事後確率を計算したかどうかを判定し、計算していれば終了し、計算していなければ、次の遺伝子に関する事後確率を計算し、

計算された各チャンネル毎の遺伝子が発現している事後確率は、前記出力装置へ送出され、

該出力装置は、各チャンネル毎の遺伝子が発現している事後確率を表示する

ことを特徴とする請求項６又は７記載の遺伝子発現状況推定システム。

【請求項９】

遺伝子発現強度データに基づいてチャンネル毎に遺伝子が発現している確率を推定する遺伝子発現状況推定方法において、

前記遺伝子発現強度データを用いて、以下の数１３に示される混合正規分布（尚、外１１は平均０，分散σ^2の１次元正規分布の密度関数を表し、外１２－１及び外１２－２はそれぞれ第１および第２のコンポーネントの平均と分散パラメータを表し，ξは混合率を表し、外１３が満たされているものとする。）の分布パラメータを推定し、この推定された分布パラメータを送出するステップと、

【数１３】

$$(1-\xi)\phi\left(u-\mu_0 \mid \sigma_0^2\right)+\xi\phi\left(u-\mu_1 \mid \sigma_1^2\right)$$

[外１１]

$$\phi\left(* \mid \sigma^2\right)$$

[外１２－１]

$$\left(\mu_0,\ \sigma_0^2\right)$$

[外１２－２]

$$\left(\mu_1,\ \sigma_1^2\right)$$

[外１３]

$$\mu_0 < \mu_1,\ \sigma_0^2 > 0,\ \sigma_1^2 > 0,\ 0 < \xi < 1$$

前記遺伝子発現強度データと、前記推定された分布パラメータとを用いて、混合正規分布の混合比パラメータを推定し、この推定された混合比パラメータを送出するステップと、

前記遺伝子発現強度データと、前記推定された分布パラメータ及び前記推定された混合比パラメータを用いて、各チャンネル毎に各遺伝子の発現状況に関する事後確率を計算し、計算された事後確率を送出するステップと、

計算された前記事後確率を出力するステップ

を有することを特徴とする遺伝子発現状況推定方法。

【請求項１０】

前記分布パラメータを推定するステップはさらに、２つのチャンネルの遺伝子発現強度Ｘ，Ｙの差を示すＶが０近傍である遺伝子発現強度データに対して、２つのコンポーネントからなる前記混合正規分布を当てはめ、混合率（ξ）、平均（μ_0，μ_1）、及び分散外１３－１を推定するステップを有する

[外１３－１]

$$\left(\sigma_0^2, \sigma_1^2\right)$$

ことを特徴とする請求項９記載の遺伝子発現状況推定方法。

【請求項１１】

前記遺伝子に関する発現量の和のデータは、遺伝子発現強度Ｘ，Ｙの差の絶対値外１４－１の中央値をc_Mとするとき、外１４－２で示される

[外１４－１]

$$|v_i|(i = 1, \cdots, n)$$

[外１４－２]

$$\left\{u_i \,|\, |v_i| \leq c_M,\ i = 1, \cdots, n\right\}$$

ことを特徴とする請求項１０記載の遺伝子発現状況推定方法。

【請求項１２】

前記分布パラメータを推定するステップにおいて、推定された外１５を用いて、外１６を以下の数１４、数１５（尚、N_0は外１７を満たすデータのインデクッス集合とし、外１８はその要素の個数とする。）、数１６、及び数１７に従って推定する

【数１４】

$$\hat{\mu} = (\hat{\mu}_1 - \hat{\mu}_0)/2$$

【数１５】

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{2\|N_0\|}\sum_{i \in N_0} v_i^2$$

【数１６】

$$\hat{\sigma}_\beta^2 = \frac{1}{4}\hat{\sigma}_0^2 - \frac{1}{2}\hat{\sigma}_\varepsilon^2$$

【数１７】

$$\hat{\lambda} = \sqrt{\log\left(1 + \frac{\hat{\sigma}_1^2 - \hat{\sigma}_0^2}{4\hat{\mu}^2}\right)}$$

[外１５]

$$\hat{\xi},\ \hat{\mu}_0,\ \hat{\sigma}_0^2,\ \hat{\mu}_1,\ \hat{\sigma}_1^2$$

[外１６]

μ，σ_ε^2，σ_β^2，λ

［外１７］

$$i \in \{ i \mid u_i < \hat{\mu}_0 \}$$

［外１８］

$$\| N_0 \|$$

ことを特徴とする請求項１１記載の遺伝子発現状況推定方法。

【請求項１３】

前記混合比パラメータを推定するステップにおいて、前記送出された遺伝子発現強度データ｛（u_i，v_i）｜ｉ＝１，．．．，ｎ｝に対して、前記送出された外１９を用いて、以下の数１８（ただし、以下の数１９及び数２０に示される関係を満たすとする。）に示される４つのコンポーネントからなる２変量混合正規分布を当てはめ、混合比パラメータｐ＝（p_{00}，p_{10}，p_{01}，p_{11}）（p_{00}は細胞１，２において、共に遺伝子が発現していない場合における混合比パラメータを表し、p_{11}は細胞１，２において、共に遺伝子が発現している場合における混合比パラメータを表し、p_{10}は細胞１において遺伝子が発現しており、細胞２において遺伝子が発現していない場合における混合比パラメータを表し、p_{01}は細胞１において遺伝子が発現していなく、細胞２において遺伝子が発現している場合における混合比パラメータを表す）を推定する

【数１８】

$$\begin{aligned} & p_{00}g_{00}(u, v \mid \hat{\theta}) + p_{10}g_{10}(u, v \mid \hat{\theta}) + p_{01}g_{01}(u, v \mid \hat{\theta}) + p_{11}g_{11}(u, v \mid \hat{\theta}) \\ & \cong p_{00}\phi\left(u \mid 4\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\varepsilon^2\right)\phi\left(v \mid 2\hat{\sigma}_\varepsilon^2\right) + p_{10}\phi_2(u - \hat{\mu}, v - \hat{\mu} \mid \Sigma_{10}) \\ & \quad + p_{01}\phi_2(u - \hat{\mu}, v + \hat{\mu} \mid \Sigma_{01}) + p_{11}\phi\left(u - 2\hat{\mu} \mid 4\hat{\mu}^2(e^{\hat{\lambda}^2} - 1)\right. \\ & \quad \left. + 4\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\varepsilon^2\right)\phi\left(v \mid 2\hat{\sigma}_\varepsilon^2\right) \end{aligned}$$

【数１９】

$$\hat{\Sigma}_{10} = \begin{pmatrix} \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 4\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\varepsilon^2 & \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) \\ \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) & \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 2\hat{\sigma}_\varepsilon^2 \end{pmatrix}$$

【数２０】

$$\hat{\Sigma}_{01}=\begin{pmatrix} \hat{\mu}^2(e^{\hat{\lambda}^2}-1)+4\hat{\sigma}_\beta^2+2\hat{\sigma}_\varepsilon^2 & -\hat{\mu}^2(e^{\hat{\lambda}^2}-1) \\ -\hat{\mu}^2(e^{\hat{\lambda}^2}-1) & \hat{\mu}^2(e^{\hat{\lambda}^2}-1)+2\hat{\sigma}_\varepsilon^2 \end{pmatrix}$$

[外１９]

$$\hat{\theta}=\left(\hat{\mu},\ \hat{\lambda},\ \hat{\sigma}_\varepsilon^2,\ \hat{\sigma}_\beta^2\right)$$

ことを特徴とする請求項１２記載の遺伝子発現状況推定方法。

【請求項１４】

前記事後確率を計算するステップでは、前記送出された遺伝子発現強度データの対（ｕ，ｖ）に対して、細胞１および細胞２における任意の遺伝子が発現している事後確率を、以下の数２１（尚、外１９－１は、以下の数２２によって与えられるものとする。）及び数２３（尚、τ_1，τ_2はそれぞれの細胞における遺伝子の真の発現の有無を１および０で表したものである。）にしたがって算出する

【数２１】

$$\Pr(\tau_1=1\mid\hat{p},\hat{\theta})=\frac{\hat{p}_{10}g_{10}(u,v\mid\hat{\theta})+\hat{p}_{11}g_{11}(u,v\mid\hat{\theta})}{f(u,v\mid\hat{p},\hat{\theta})}$$

【数２２】

$$f(u,v\mid\hat{p},\hat{\theta})=\sum_{(j,k)\in\{0,1\}^2}\hat{p}_{jk}g_{jk}(u,v\mid\hat{\theta})$$

【数２３】

$$\Pr(\tau_2=1\mid\hat{p},\hat{\theta})=\frac{\hat{p}_{01}g_{01}(u,v\mid\hat{\theta})+\hat{p}_{11}g_{11}(u,v\mid\hat{\theta})}{f(u,v\mid\hat{p},\hat{\theta})}$$

[外１９－１]

$$f\left(u,v\mid\hat{p},\hat{\theta}\right)$$

ことを特徴とする請求項１３記載の遺伝子発現状況推定方法。

【請求項１５】

前記事後確率を計算するステップでは、細胞１および細胞２で遺伝子の発現状態が異なる状態である事後確率を、以下の数２４（尚、τ_1，τ_2はそれぞれの細胞における遺伝子の真の発現の有無を１および０で表したものである。）にしたがって算出する

【数２４】

$$Pr(\tau_1 \neq \tau_2 \mid \hat{p}, \hat{\theta}) = \frac{\hat{p}_{10} g_{10}(u, v \mid \hat{\theta}) + \hat{p}_{01} g_{01}(u, v \mid \hat{\theta})}{f(u, v \mid \hat{p}, \hat{\theta})}$$

ことを特徴とする請求項１３記載の遺伝子発現状況推定方法。

【請求項１６】

前記事後確率を計算するステップでは、全ての遺伝子発現強度データの対（ｕ，ｖ）に基づいて、遺伝子が発現している事後確率を計算したかどうかを判定し、計算していれば終了し、計算していなければ、次の遺伝子に関する事後確率を計算する

ことを特徴とする請求項１４又は１５記載の遺伝子発現状況推定方法。

【請求項１７】

遺伝子発現強度データに基づいてチャンネル毎に遺伝子が発現している確率を推定するためコンピュータに、

前記遺伝子発現強度データを用いて、以下の数２５に示される混合正規分布（尚、外２０は平均０，分散σ^2の１次元正規分布の密度関数を表し、外２１－１及び外２１－２はそれぞれ第１および第２のコンポーネントの平均と分散パラメータを表し，ξは混合率を表し、外２２が満たされているものとする。）の分布パラメータを推定し、この推定された分布パラメータを送出するステップと、

【数２５】

$$(1-\xi)\phi\left(u-\mu_0 \mid \sigma_0^2\right)+\xi\phi\left(u-\mu_1 \mid \sigma_1^2\right)$$

[外２０]

$$\phi\left(* \mid \sigma^2\right)$$

[外２１－１]

$$\left(\mu_0, \sigma_0^2\right)$$

[外２１－２]

$$\left(\mu_1, \sigma_1^2\right)$$

[外２２]

$$\mu_0 < \mu_1, \sigma_0^2 > 0, \sigma_1^2 > 0, 0 < \xi < 1$$

前記遺伝子発現強度データと、前記推定された分布パラメータとを用いて、混合正規分布の混合比パラメータを推定し、この推定された混合比パラメータを送出するステップと、

前記遺伝子発現強度データと、前記推定された分布パラメータ及び前記推定された混合比パラメータを用いて、各チャンネル毎に各遺伝子の発現状況に関する事後確率を計算し、計算された事後確率を送出するステップと、

計算された前記事後確率を出力するステップ

を実行させるための遺伝子発現状況推定プログラム。

【請求項１８】
遺伝子発現強度データに基づいてチャンネル毎に遺伝子が発現している確率を推定するためコンピュータに、
前記遺伝子発現強度データを用いて、以下の数２６に示される混合正規分布（尚、外２３は平均０，分散σ^2の１次元正規分布の密度関数を表し、外２４－１及び外２４－２はそれぞれ第１および第２のコンポーネントの平均と分散パラメータを表し，ξは混合率を表し、外２５が満たされているものとする。）の分布パラメータを推定し、この推定された分布パラメータを送出するステップと、
【数２６】

$$(1-\xi)\phi\left(u-\mu_0 \mid \sigma_0^2\right)+\xi\phi\left(u-\mu_1 \mid \sigma_1^2\right)$$

[外２３]

$$\phi\left(* \mid \sigma^2\right)$$

[外２４－１]

$$\left(\mu_0,\ \sigma_0^2\right)$$

[外２４－２]

$$\left(\mu_1,\ \sigma_1^2\right)$$

[外２５]

$$\mu_0 < \mu_1,\ \sigma_0^2 > 0,\ \sigma_1^2 > 0,\ 0 < \xi < 1$$

前記遺伝子発現強度データと、前記推定された分布パラメータとを用いて、混合正規分布の混合比パラメータを推定し、この推定された混合比パラメータを送出するステップと、
前記遺伝子発現強度データと、前記推定された分布パラメータ及び前記推定された混合比パラメータを用いて、各チャンネル毎に各遺伝子の発現状況に関する事後確率を計算し、計算された事後確率を送出するステップと、
計算された前記事後確率を出力するステップ
を実行させるための遺伝子発現状況推定プログラムを記録したコンピュータ読み取り可能な記録媒体。

【書類名】明細書
【発明の名称】遺伝子発現状況推定システム、方法、プログラム、及び記録媒体
【技術分野】
【０００１】
本発明は、二色蛍光法によるｃＤＮＡマイクロアレイデータの統計的解析方法、解析システム、及び記録媒体に関し、特にチャンネル毎に遺伝子が発現している確率を推定するシステム、方法、プログラム、及び記録媒体に関するものである。
【背景技術】
【０００２】
現在、ゲノム研究は個々の遺伝子についての構造解析から体系的な遺伝子の機能解析へと展開しつつある。機能未知の遺伝子や総体としての遺伝子の機能解析のために、多数の遺伝子の発現強度を同時に定量化することのできるｃＤＮＡ（相補的なＤＮＡ）マイクロアレイを用いた実験はその有効性が大いに期待されている。
【０００３】
二色蛍光法によるｃＤＮＡマイクロアレイを用いた実験の目的は二種類の細胞の遺伝子発現強度の違いを検出することにある。ここで、二色蛍光法によるｃＤＮＡマイクロアレイの概要について述べる。まず、多数の遺伝子セットのｃＤＮＡを参照用のプローブとして、スライドグラス上にアレイ状に高密度に固定化する（マイクロアレイ）。
【０００４】
次に、条件の異なる２種類のサンプル、細胞１と細胞２（例えば正常細胞と癌細胞）から抽出したｍＲＮＡをそれぞれ波長の異なる蛍光色素でラベルし、ターゲットｃＤＮＡを合成する。そして、それらを等量混合したものをマイクロアレイに固定化された参照用のプローブｃＤＮＡに競合的にハイブリダイズさせる。ハイブリダイゼーション後、スキャナーでそれぞれの蛍光色素強度を測定する。細胞１にラベルされた蛍光色素をチャンネル1により、細胞２にラベルされた蛍光色素をチャンネル２により読み取り、それぞれを各細胞の遺伝子発現強度データ（マイクロアレイデータ）とする。
【０００５】
このように、マイクロアレイデータが得られるまでの過程は複雑であり、高度な実験技術が必要とされることから、実験の各段階において様々な実験誤差が生じると考えられる。このため、マイクロアレイデータから真に生物学的意味のあるデータを取り出すためには発現強度の分布と実験誤差の解析は解決すべき重要な課題である。
【０００６】
発現強度の分布に関しては、例えば、以下の非特許文献１を参照すると、Ｎｅｗｔｏｎ等は発現強度にガンマ分布関数を仮定し、発現強度比（チャンネル１とチャンネル２の発現強度の比）についての統計学的性質を考察している。
【０００７】
また、観測された発現強度データに対しては、例えば、以下の非特許文献２を参照すると、Ｌｅｅ等は真の発現強度を２個の水準値に分離できることおよび偶然誤差の存在を前提として、以下の数２７に示されるような混合正規分布を適用し、発現強度データについての統計学的考察を行っている。
【０００８】
【数２７】

$$f(x) = p\phi\left(x - \mu_1 \mid \sigma_1^2\right) + (1 - p)\phi\left(x - \mu_2 \mid \sigma_2^2\right)$$

ここで、ｘはスキャナーなどによって得られる蛍光強度などの遺伝子発現強度（の対数値）を表し、右辺第１項の外２６は遺伝子が発現しているときの平均μ_1，分散外２７の正規分布、

[外２６]

$$\phi\left(x-\mu_1 \mid \sigma_1^2\right)$$

[外２７]

$$\sigma_1^2$$

また、同第２項の外２８は遺伝子が発現していないときの平均μ_2、分散外２９の正規分布の密度関数を表し，ｐはその混合率を表す母数である。

【０００９】

[外２８]

$$\phi\left(x-\mu_2 \mid \sigma_2^2\right)$$

[外２９]

$$\sigma_2^2$$

【００１０】

実験誤差の解析については、系統誤差の除去、いわゆるノーマライゼーションの方法がいくつか提案されている。非特許文献３を参照すると、Ｃｈｅｎ等は二つの細胞の遺伝子発現強度の中央値は等しいとしてチャンネル１とチャンネル２で得られた測定値の補正を行っている。また，非特許文献４、５、６を参照すると、ＤｕｄｏｉｔやＳｃｈｕｃｈｈａｒｄｔやＹａｎｇは、系統誤差がスポットのスライドグラス上の位置や、二種類の蛍光色素の感度の違いによって生じたものと考え、それらを除去する方法を提案している。

【非特許文献１】Newton et. al、2001年、ジャーナル・オブ・コンピュテーショナル・バイオロジー、第8巻、37～52頁（Journal of Computational Biology Vol. 8, pp. 37-52）

【非特許文献２】Lee et. al、2000年、プロシーディング・オブ・ザ・ナショナル・アカデミィー・オブ・サイエンシズ、第97巻、第18号、9834～9839頁（Proceeding of the National Academy of Sciences Vol. 97, No 18, pp. 9834-9839）

【非特許文献３】Chen et. al、1997年、ジャーナル・オブ・バイオメディカル・オプティクス、第2号、364～374頁（Journal of Biomedical Optics Vol. 2, pp. 364-374）

【非特許文献４】Dudoit et. al、2000. Statistical methods for identifying differentially expressed genes in replicated ｃＤＮＡmicroarray experiments. Technical Report #578 2.

【非特許文献５】Schuchhardt et. al、2000年、ヌクレ・アシッド・リサーチ、第28巻、第10号（Nucleic Acids Research, 2000, Vol.28, No. 10）

【非特許文献６】Yang et. al、 2002年、ヌクレ・アシッド・リサーチ、第30巻、第4号（Nucleic Acids Research, 2002, Vol.30, No. 4）

【発明の開示】

【発明が解決しようとする課題】

【００１１】

上記した従来技術における問題点は、マイクロアレイデータの解析結果は再現性に乏しく解析の精度や効率が低いことにある。その原因として，従来の解析方法では，得られたマイクロアレイデータが真の信号と系統誤差および測定誤差に十分に分離できていないことが考えられる。したがって，系統誤差の除去および測定誤差の評価が重要な課題となる

。

【００１２】

系統誤差の除去に関しては「ｃＤＮＡ マイクロアレイデータの補正法、システム、及び記録媒体」として，別途特許にて申請中である。本発明におけるマイクロアレイデータは、系統誤差はあらかじめ取り除かれているものとする。

【００１３】

従来のマイクロアレイデータを用いた解析では，各スポットのｃＤＮＡ量の均一性が保証されていないという理由で、２つのチャンネルの遺伝子発現強度の比（対数値の差）のみを扱っており、各チャンネル毎の遺伝子発現強度を扱っていない。従って、遺伝子の発現状況に関する真の信号と測定誤差の分離が十分に行われていない。

【００１４】

本発明の目的は、マイクロアレイデータを用いた解析の精度および効率を高めるために、遺伝子の発現に関する真の信号と測定誤差の分離を行い、さらに，各チャンネル毎に、遺伝子が発現している確率を推定する手法およびシステムを提供することにある。

【課題を解決するための手段】

【００１５】

本発明の遺伝子発現状況推定システムは、マイクロアレイデータを入力する入力装置と、プログラム制御により動作するデータ解析装置と、出力装置とを含む。前記データ解析装置は、前記入力装置から与えられた遺伝子発現強度データを用いて、混合正規分布のコンポーネント毎の分布パラメータおよび混合比パラメータを推定するパラメータ推定手段と、推定された各パラメータを用いて、各チャンネル毎に遺伝子の発現に関する事後確率を計算する事後確率計算手段を有し、計算された事後確率は前記出力装置に出力することを特徴とする。

【００１６】

このような構成を採用し、遺伝子発現状況を推定することにより、本発明の目的は達成することができる。

【発明の効果】

【００１７】

本発明による第１の効果は、マイクロアレイによって得られた遺伝子発現強度データに対して遺伝子の発現および非発現という概念を導入し，数理モデルを構築することにより遺伝子の発現に関する真の信号と実験誤差との分離が可能となったことにある。

【００１８】

本発明による第２の効果は、マイクロアレイによって得られた遺伝子発現強度データを、２つのチャンネルの遺伝子発現強度データの和と差のデータに変換することにより，２つのチャンネルの蛍光強度の感度情報が得られ易くなったことにある。この結果、感度の違いによる実験誤差の大きさを視覚的に示すことが可能となった。

【００１９】

本発明による第３の効果は、マイクロアレイによって得られた遺伝子発現強度データを、２つのチャンネルの遺伝子発現強度の和と差のデータに変換し，それらの２次元同時分布を記述することにより，２つのチャンネル毎に、各遺伝子の発現および非発現に関する事後確率を推定することが可能となったことにある。その結果、細胞１と細胞２の違いに関係する遺伝子を高い精度で検出することが可能となった。

【発明を実施するための最良の形態】

【００２０】

はじめに本発明におけるマイクロアレイによって得られた遺伝子発現強度データに対する数理モデルを説明する。Ｘをチャンネル１によって得られた細胞１の遺伝子発現強度とし、Ｙをチャンネル２によって得られた細胞２の遺伝子発現強度としたときに、それぞれの遺伝子発現データを数２８で示す。なお，ＸおよびＹは観察値に対して対数変換ないしは適当なべき変換および１次変換を含む適当な変換が施されている量である。

【００２１】

【数２８】

$$X = \tau_1\alpha + \beta + \varepsilon_1$$

$$Y = \tau_2\alpha + \beta + \varepsilon_2$$

【００２２】
ここで、τ_1，τ_2はそれぞれの細胞における遺伝子の真の発現の有無（ＯＮ／ＯＦＦ）を１および０で表す。また、αは遺伝子がＯＮの状態にある場合に産生されるｍＲＮＡの量およびスポットの状態に規定される遺伝子発現強度変量とし、βはチャンネル１およびチャンネル２に共通な測定誤差とし、εはチャンネル間で独立な測定誤差とする。なお、それぞれの確率変数の分布は数２９に従うものとする。
【００２３】
【数２９】

$$\log\alpha \sim N\left(\mu - \frac{\lambda^2}{2}, \lambda^2\right)$$

$$\varepsilon_j \sim N\left(0, \sigma_\varepsilon^2\right), \quad j = 1,2$$

$$\beta \sim N\left(0, \sigma_\beta^2\right)$$

【００２４】
ここで、Ｎ（μ，σ^2）は平均μ，分散σ^2の１次元正規分布とする。また、α，βおよびεはそれぞれ独立とする。この数理モデルにおいては、遺伝子が発現している場合（ＯＮ状態）の真の発現強度は非負値をとる確率変数であり、遺伝子が発現していない場合（ＯＦＦ状態）は、単なる実験誤差のみが観察されるとしている。さらに、非特許文献６を参照して、Ｙａｎｇ ＹＨらが導入したＭ－Ａ変換を変形した数３０で示すようなＳ－Ｄ変換を行う。
【００２５】
【数３０】

$$U = X + Y,$$

$$V = X - Y,$$

【００２６】
すなわち、２つのチャンネルの遺伝子発現強度の和をＵとし、差をＶと変換する。このＳ－Ｄ変換を行ったときの本モデルの模式図を図１に示す。なお、以下ではこのプロットをＳ－Ｄプロットとよぶ。図１において、g_{00}はそれぞれの細胞において遺伝子が発現していない同時分布を表し、g_{10}は細胞１において遺伝子が発現しており、細胞２において遺伝子が発現していない同時分布を表し、g_{01}は細胞１において遺伝子が発現していなく、細胞２において遺伝子が発現している同時分布を表し、g_{11}はそれぞれの細胞において遺伝子が発現している同時分布を表している。g_{00}の密度関数を数３１に示す。
【００２７】

【数３１】

$$g_{00}(u, v \mid \theta) = \phi\left(u \mid 4\sigma_\beta^2 + 2\sigma_\varepsilon^2\right)\phi\left(v \mid 2\sigma_\varepsilon^2\right)$$

ここで、ϕ（u｜σ^2）は平均０分散σ^2の１次元正規分布の密度関数とする。g_{10}の密度関数を数３２に示す。

【００２８】

【数３２】

$$g_{10}(u, v \mid \theta)$$

$$= \int_{-\infty}^{+\infty} \phi\left(u - \mu e^{-\frac{\lambda^2}{2}+\lambda z} \mid 4\sigma_\beta^2 + 2\sigma_\varepsilon^2\right)\phi\left(v - \mu e^{-\frac{\lambda^2}{2}+\lambda z} \mid 2\sigma_\varepsilon^2\right)\phi(z|1)dz$$

$$\cong \phi_2(u - \mu, v - \mu \mid \Sigma_{10})$$

ここで、ϕ_2（u，v｜Σ）は平均ベクトル０、分散共分散行列Σの２次元正規分布の密度関数とし、Σ_{10}は２×２の分散共分散行列であり、数３３に示す。

【００２９】

【数３３】

$$\Sigma_{10} = \begin{pmatrix} \mu^2(e^{\lambda^2} - 1) + 4\sigma_\beta^2 + 2\sigma_\varepsilon^2 & \mu^2(e^{\lambda^2} - 1) \\ \mu^2(e^{\lambda^2} - 1) & \mu^2(e^{\lambda^2} - 1) + 2\sigma_\varepsilon^2 \end{pmatrix}$$

g_{01}の密度関数を数３４に示す。

【００３０】

【数３４】

$$g_{01}(u, v \mid \theta)$$

$$= \int_{-\infty}^{+\infty} \phi\left(u - \mu e^{-\frac{\lambda^2}{2}+\lambda z} \mid 4\sigma_\beta^2 + 2\sigma_\varepsilon^2\right)\phi\left(v + \mu e^{-\frac{\lambda^2}{2}+\lambda z} \mid 2\sigma_\varepsilon^2\right)\phi(z|1)dz$$

$$\cong \phi_2(u - \mu, v + \mu \mid \Sigma_{01})$$

ここで、Σ_{01}は２×２の分散共分散行列であり、数３５に示す。

【００３１】

【数３５】

$$\Sigma_{01}=\begin{pmatrix} \mu^2(e^{\lambda^2}-1)+4\sigma_\beta^2+2\sigma_\varepsilon^2 & -\mu^2(e^{\lambda^2}-1) \\ -\mu^2(e^{\lambda^2}-1) & \mu^2(e^{\lambda^2}-1)+2\sigma_\varepsilon^2 \end{pmatrix}$$

g_{11}の密度関数を数３６に示す。

【００３２】

【数３６】

$$g_{11}(u,v\mid\theta)=\phi\left(v\mid 2\sigma_\varepsilon^2\right)\int_{-\infty}^{+\infty}\phi\left(u-\mu e^{-\frac{\lambda^2}{2}+\lambda z}\mid 4\sigma_\beta^2+2\sigma_\varepsilon^2\right)\phi(z\mid 1)dz$$

$$\cong\phi\left(u-2\mu\mid 4\mu^2(e^{\lambda^2}-1)+4\sigma_\beta^2+2\sigma_\varepsilon^2\right)\phi\left(v\mid 2\sigma_\varepsilon^2\right)$$

以上の分布をもとに、細胞１および細胞２における遺伝子が発現している事後確率を数３７および数３８に示す。

【００３３】

【数３７】

$$Pr(\tau_1=1\mid p,\theta)=\frac{p_{10}g_{10}(u,v\mid\theta)+p_{11}g_{11}(u,v\mid\theta)}{f(u,v\mid p,\theta)}$$

【数３８】

$$Pr(\tau_2=1\mid p,\theta)=\frac{p_{01}g_{01}(u,v\mid\theta)+p_{11}g_{11}(u,v\mid\theta)}{f(u,v\mid p,\theta)}$$

ここで、ｆ（ｕ，ｖ｜ｐ，θ）は数３９によって与えられるものとする。

【００３４】

【数３９】

$$f(u,v\mid p,\theta)=\sum_{(j,k)\in\{0,1\}^2}p_{jk}g_{jk}(u,v\mid\theta)$$

ただし、ただし、ｐ＝（p_{00}，p_{10}，p_{01}，p_{11}）は各分布の混合率を表すパラメータとする。

【００３５】

次に、本発明の実施の形態について図面を参照して詳細に説明する。図２を参照すると、本発明の第１の実施の形態は、遺伝子発現強度データに関する数理モデルの定式化およびそのデータ解析への適用による未知母数の推定する過程及び算出された母数の推定値を

使用して、細胞１と細胞２の遺伝子の発現状態に関する事後推定を行うためのシステムである。同システムにはキーボード等の入力装置１と、プログラムの制御により動作するデータ解析装置２と、ディスプレイ装置や印刷装置等の出力装置３が含まれる。

【００３６】

データ解析装置２は、分布パラメータ推定手段２１と、混合比パラメータ推定手段２２と、事後確率計算手段２３とを備えている。分布パラメータ推定手段２１は、入力装置１から与えられた遺伝子発現強度データを用いて、混合正規分布の各コンポーネント毎の分布パラメータを推定する。推定された分布パラメータは、混合比パラメータ推定手段２２と、事後確率計算手段２３へ送られる。混合比パラメータ推定手段２２は、入力装置１から与えられた遺伝子発現強度データと、分布パラメータ推定手段２１から与えられた各コンポーネント毎の分布パラメータを用いて、混合正規分布の混合比パラメータを条件付き最尤法により推定する。推定された混合比パラメータは、事後確率計算手段２３へ送られる。事後確率計算手段２３は、入力装置１から与えられた遺伝子発現強度データと、分布パラメータ推定手段２１から与えられた各コンポーネント毎の分布パラメータと、混合比パラメータ推定手段２２から与えられた混合比パラメータを用いて、各チャンネル毎に遺伝子の発現状況に関する事後確率を計算する。計算された事後確率は出力装置３へ送られる。

【００３７】

次に、図２及び図３を参照して、遺伝子発現強度データに関する数理モデルの定式化およびそのデータ解析への適用による未知母数の推定する過程について詳細に説明する。入力装置１から与えられた遺伝子発現強度データ外３０は、分布パラメータ推定手段２１及び混合比パラメータ推定手段２２へ送られる。分布パラメータ推定手段２１は遺伝子発現強度の差の絶対値外３１の中央値をc_Mとするとき、Ｖ＝０近傍の遺伝子に関する発現量の和のデータ外３２に対して、以下の数４０に示す、２つのコンポーネントからなる混合正規分布を当てはめ、ξ，μ_0，σ_0，μ_1，σ_1を推定する（図３のステップＡ１）。

【００３８】

【数４０】

$$(1-\xi)\phi\left(u-\mu_0 \mid \sigma_0^2\right)+\xi\phi\left(u-\mu_1 \mid \sigma_1^2\right)$$

[外３０]

$$\left\{(u_i, v_i) \mid i=1,\cdots,n\right\}$$

[外３１]

$$\left|v_i\right| (i=1,\cdots,n)$$

[外３２]

$$\left\{u_i \mid \left|v_i\right| \leq c_M,\ i=1,\cdots,n\right\}$$

【００３９】

ここで、外３３は、平均０分散σ^2の１次元正規分布の密度関数であり、外３４と外３５はそれぞれ第１および第２コンポーネントの平均と分散パラメータ、ξは混合率であり、外３６が満たされているものとする。

【００４０】

次に、分布パラメータ推定手段２１は、推定された外３７を用いて、外３８を以下の数

４１、数４２、数４３、及び数４４に従って推定する（ステップＡ２）。

【００４１】

[外３３]

$$\phi\left(*\mid\sigma^2\right)$$

[外３４]

$$\left(\mu_0,\ \sigma_0^2\right)$$

[外３５]

$$\left(\mu_1,\ \sigma_1^2\right)$$

[外３６]

$$\mu_0<\mu_1,\ \sigma_0^2>0,\ \sigma_1^2>0,\ 0<\xi<1$$

[外３７]

$$\hat{\xi},\ \hat{\mu}_0,\ \hat{\sigma}_0^2,\ \hat{\mu}_1,\ \hat{\sigma}_1^2$$

[外３８]

$$\mu,\ \sigma_\varepsilon^2,\ \sigma_\beta^2,\ \lambda$$

【数４１】

$$\hat{\mu}=(\hat{\mu}_1-\hat{\mu}_0)/2$$

【数４２】

$$\hat{\sigma}_\varepsilon^2=\frac{1}{2\|N_0\|}\sum_{i\in N_0}v_i^2$$

【数４３】

$$\hat{\sigma}_\beta^2=\frac{1}{4}\hat{\sigma}_0^2-\frac{1}{2}\hat{\sigma}_\varepsilon^2$$

【数４４】

$$\hat{\lambda} = \sqrt{\log\left(1 + \frac{\hat{\sigma}_1^2 - \hat{\sigma}_0^2}{4\hat{\mu}^2}\right)}$$

【００４２】

ここで、N_0は外３９を満たすデータのインデックス集合とし、外４０はその要素の個数とする。

【００４３】

[外３９]

$$i \in \{ i \mid u_i < \hat{\mu}_0 \}$$

[外４０]

$$\| N_0 \|$$

【００４４】

次に、混合比パラメータ推定手段２２は、入力装置１から与えられた遺伝子発現強度データ｛$(u_i, v_i) \mid i = 1, \ldots, n$｝に対して、分布パラメータ推定手段２１から与えられた各パラメータの推定値外４１を用いて、以下の数４５（尚、以下の数４６（ただし、外４２は数３３から導かれる２×２の分散共分散行列）及び数４７（ただし、外４３は数３５から導かれる２×２の分散共分散行列）に示される関係を満たすとする。）に示される４つのコンポーネントからなる２変量混合正規分布を当てはめ、混合比パラメータ$p = (p_{00}, p_{10}, p_{01}, p_{11})$を条件付き最尤法により推定する（ステップＡ３）。

【００４５】

[外４１]

$$\hat{\theta} = \left(\hat{\mu}, \hat{\lambda}, \hat{\sigma}_\varepsilon^2, \hat{\sigma}_\beta^2\right)$$

【数４５】

$$p_{00}g_{00}(u, v \mid \hat{\theta}) + p_{10}g_{10}(u, v \mid \hat{\theta}) + p_{01}g_{01}(u, v \mid \hat{\theta}) + p_{11}g_{11}(u, v \mid \hat{\theta})$$

$$\cong p_{00}\phi\left(u \mid 4\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\varepsilon^2\right)\phi\left(v \mid 2\hat{\sigma}_\varepsilon^2\right) + p_{10}\phi_2(u - \hat{\mu}, v - \hat{\mu} \mid \Sigma_{10})$$

$$+ p_{01}\phi_2(u - \hat{\mu}, v + \hat{\mu} \mid \Sigma_{01}) + p_{11}\phi\left(u - 2\hat{\mu} \mid 4\hat{\mu}^2(e^{\hat{\lambda}^2} - 1)\right.$$

$$\left. + 4\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\varepsilon^2\right)\phi\left(v \mid 2\hat{\sigma}_\varepsilon^2\right)$$

【数４６】

$$\hat{\Sigma}_{10}=\begin{pmatrix}\hat{\mu}^2(e^{\hat{\lambda}^2}-1)+4\hat{\sigma}_\beta^2+2\hat{\sigma}_\varepsilon^2 & \hat{\mu}^2(e^{\hat{\lambda}^2}-1)\\ \hat{\mu}^2(e^{\hat{\lambda}^2}-1) & \hat{\mu}^2(e^{\hat{\lambda}^2}-1)+2\hat{\sigma}_\varepsilon^2\end{pmatrix}$$

【数４７】

$$\hat{\Sigma}_{01}=\begin{pmatrix}\hat{\mu}^2(e^{\hat{\lambda}^2}-1)+4\hat{\sigma}_\beta^2+2\hat{\sigma}_\varepsilon^2 & -\hat{\mu}^2(e^{\hat{\lambda}^2}-1)\\ -\hat{\mu}^2(e^{\hat{\lambda}^2}-1) & \hat{\mu}^2(e^{\hat{\lambda}^2}-1)+2\hat{\sigma}_\varepsilon^2\end{pmatrix}$$

[外４２]

$\hat{\Sigma}_{10}$

[外４３]

$\hat{\Sigma}_{01}$

【００４６】
次に、算出された母数の推定値を使用して、細胞１と細胞２の遺伝子の発現状態に関する事後推定を行う過程を説明する。
【００４７】
各細胞の遺伝子毎の発現状況に関する事後確率計算手段２３は、入力装置１から与えられた遺伝子発現強度データの対（ｕ，ｖ）に対して、分布パラメータ推定手段２１及び混合比パラメータ推定手段２２から与えられた各パラメータの推定値外４４および外４５を用いて記述できる。
【００４８】
[外４４]

$$\hat{\theta}=\left(\hat{\mu},\ \hat{\lambda},\ \hat{\sigma}_\varepsilon^2,\ \hat{\sigma}_\beta^2\right)$$

[外４５]

$$\hat{p}=\left(\hat{p}_{00},\ \hat{p}_{10},\ \hat{p}_{01},\ \hat{p}_{11}\right)$$

【００４９】
すなわち、細胞１および細胞２における任意の遺伝子の発現がＯＮになっている事後確率は、以下の数４８及び数４９より算出できる（ステップＡ４）。
【００５０】

【数４８】

$$Pr(\tau_1 = 1 \mid \hat{p}, \hat{\theta}) = \frac{\hat{p}_{10} g_{10}(u, v \mid \hat{\theta}) + \hat{p}_{11} g_{11}(u, v \mid \hat{\theta})}{f(u, v \mid \hat{p}, \hat{\theta})}$$

【数４９】

$$Pr(\tau_2 = 1 \mid \hat{p}, \hat{\theta}) = \frac{\hat{p}_{01} g_{01}(u, v \mid \hat{\theta}) + \hat{p}_{11} g_{11}(u, v \mid \hat{\theta})}{f(u, v \mid \hat{p}, \hat{\theta})}$$

【００５１】
全ての遺伝子発現強度データの対（ｕ，ｖ）に基づいて、遺伝子の発現がＯＮになっている事後確率を計算したかどうかを判定し（ステップＡ５）、計算していれば終了する。計算していなければ、次の遺伝子に関する事後確率を計算する。

【００５２】
計算された各チャンネル毎の遺伝子が発現している事後確率は、出力装置３へ送る。出力装置３は、各チャンネル毎の遺伝子が発現している事後確率をグラフで表示したり、印刷したりする。

【００５３】
次に、本実施の形態の効果について説明する。本実施の形態では、マイクロアレイで得られた遺伝子発現データに対して、遺伝子の発現および非発現を導入した数理モデルを構築し、真の信号と実験誤差との分離を行った。また、２つのチャンネルの遺伝子発現強度による和と差のデータを用いることにより、マイクロアレイデータでの各チャンネルの蛍光強度の感度情報が得られ易くなり、実験誤差の大きさをより的確に抽出することが可能となった。さらに、これらの和と差のデータに対して２次元同時分布を記述することにより、チャンネル毎に各遺伝子の発現に関して、高い精度の事後確率を推定することが可能となった。

【００５４】
また、細胞１と細胞２で遺伝子の発現状態が異なるという事象（ＯＮ－ＯＦＦが不一致である状態）（図３のステップＡ４）の事後確率は以下の数５０により算出される。

【００５５】
【数５０】

$$Pr(\tau_1 \neq \tau_2 \mid \hat{p}, \hat{\theta}) = \frac{\hat{p}_{10} g_{10}(u, v \mid \hat{\theta}) + \hat{p}_{01} g_{01}(u, v \mid \hat{\theta})}{f(u, v \mid \hat{p}, \hat{\theta})}$$

【００５６】
本実施の形態の効果として、細胞１と細胞２において発現状態が異なる可能性の高い遺伝子の候補を検出することが可能となる。

【００５７】
次に、本発明の第２の実施の形態について図面を参照して詳細に説明する。

【００５８】
図４を参照すると、本発明の第２の実施形態は、本発明の第１の実施形態と同様に、入力装置、データ解析装置、出力装置を備え、更に、データ解析プログラムを記録した記録媒体４を備える。この記録媒体４は可搬形あるいは固定型のいずれであってもよく、磁気ディスク、半導体メモリ、ＣＤ－ＲＯＭその他の記録媒体であってもよい。また、本手法を実行できるコンピュータプログラムを、ネットワークに接続されたコンピュータの記録装置に格納しておき、ネットワークを介して他のコンピュータに転送することもできる。本アルゴリズムを実行するコンピュータプログラムを提供する提供媒体としては、様々な形

式のコンピュータに読み出し可能な媒体として頒布可能であって、特定のタイプの媒体に限定されるものではない。

【００５９】

データ解析プログラムは記録媒体４からデータ解析装置５に読み込まれ、データ解析装置５の動作を制御し、入力装置１から入力されたデータファイルに対して第１の実施の形態におけるデータ処理装置２による処理と同一の処理を実行する。

【実施例１】

【００６０】

次に、本発明の実施例について説明する。例として用いたデータは，異なる２種類の癌細胞（細胞１、細胞２）の遺伝子発現状況の比較のために行われた実験から得られたものである。

【００６１】

一枚のチップ上に４８グリッド，１グリッドあたり４４１（２１×２１）スポット，計２１１６８の遺伝子の発現パターンについて調べたものである。

【００６２】

図５および図６に、細胞１および細胞２の遺伝子の発現状態が両者ともにＯＦＦ、あるいは両者ともにＯＮである場合（Ｖ＝０）、それらの発現強度の和の分布Ｕに混合正規分布と、その対比として単峰の正規分布を当てはめた場合の各分布の推定結果を示す。以下の表１にコンポーネント毎の分布パラメータの推定結果を示す（図３のステップＡ１の結果）。

【００６３】

【表１】

Result of N2MIXFit (Ver 0.998)

Name of Data Set to be analyzed = n145h1.std
Name of Target Variate = S_CH1
Type of Transformation = 1/16
Sample size = 14726
Critical Value for Convergence = .10000E-06
Iterations for Convergence = 50
Job Termination Status = Normally Terminated

	Mean	SD	Rate(%)
Single Component:	3.0273	4.3560	100.00
1st Component:	.89956	1.2978	47.38
2nd Component:	4.9430	5.1395	52.62

Log_Likelihood for Single Component = -42565.
for Two Components Mixed = -40465.
Log of Likelihood Ratio Statistics = 2099.8

【００６４】

図５に推定された累積分布関数を、図６に推定された密度関数を示す。細実線は混合正規分布を仮定した場合、二点鎖線はその第１コンポーネント（ＯＦＦ－ＯＦＦ）、太実線は第２コンポーネント（ＯＮ－ＯＮ）、そして点線は単峰の正規分布を仮定した場合の推定結果を示している。

【００６５】

図５における一点鎖線は観察されたデータに基づく経験累積分布関数を示しており、観察値が推定された混合正規分布（細実線）によく従っていることを示している。

【００６６】

図６において＊印（尚、左右両端は＊印が確認できるが、遺伝子発現強度が０付近から

３０付近までは重なり合っているため＊印ではなく以下の（１）～（５）に示すハッチングで表示することとした。）は観察データを示しており、それぞれのハッチング領域（（１）～（５））は、第１コンポーネントに属している事後確率の大きさを示している。（１）の黒塗り領域は０～０．２、（２）のハッチング領域は０．２～０．４、（３）のハッチング領域は０．４～０．６、（４）のハッチング領域は０．６～０．８、（５）のハッチング領域は０．８～１．０を表している。

【００６７】

図７に、遺伝子発現強度データのＳ－Ｄプロットを示す。横軸は細胞１と細胞２の遺伝子発現強度の対数値の和を、縦軸はそれらの差を示している。各マークの色は、細胞１と細胞２の遺伝子の発現状態が不一致（ＯＮ－ＯＦＦあるいはＯＦＦ－ＯＮ）である事後確率の大きさを示している。（１）の黒塗り領域は０～０．２、（２）のハッチング領域は０．２～０．４、（３）のハッチング領域は０．４～０．６、（４）のハッチング領域は０．６～０．８、（５）のハッチング領域は０．８～１．０を表している。

【００６８】

以下の表２に、条件付き最尤法による混合比パラメータの推定結果を示す（図３のステップＡ２，Ａ３の結果）。図７における各プロットに対しては、それぞれ遺伝子が対応しているので、細胞１と細胞２の違いに関係する遺伝子の候補を容易に絞り込むことができる。

【００６９】

【表２】

RESULT OF MAD Ver(0.998)

==========

Transformation = 6　(1/16)

Samplre Size =21168

SIGMA_Epsilon =	.898
MU =	2.022
LAMBDA =	.960
SIGMA_Beta =	.133
P11 =	.561
P10 =	.004
P01 =	.011
P00 =	.425

【図面の簡単な説明】

【００７０】

【図１】本発明における数理モデルのＳ－Ｄプロットによる模式図である。

【図２】本発明の第１の実施の形態の構成を示すブロック図である。

【図３】本発明の第１の実施の形態の動作を示すフローチャートである。

【図４】本発明の第２の実施形態の構成を示すブロック図である。

【図５】Ｖ＝０近傍の遺伝子発現強度データに対して推定した正規分布の累積分布グラフである。

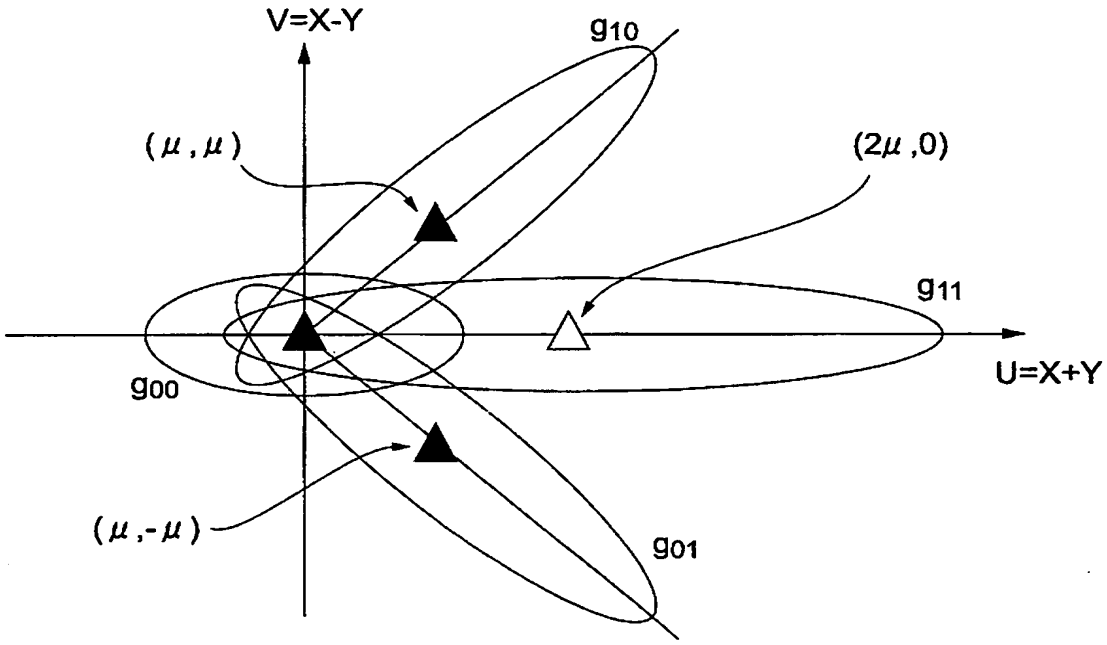
【図６】Ｖ＝０近傍の遺伝子発現強度データに対して推定した正規分布の密度関数を示したグラフである。

【図７】遺伝子発現強度データのＳ－Ｄプロットを示した図である。

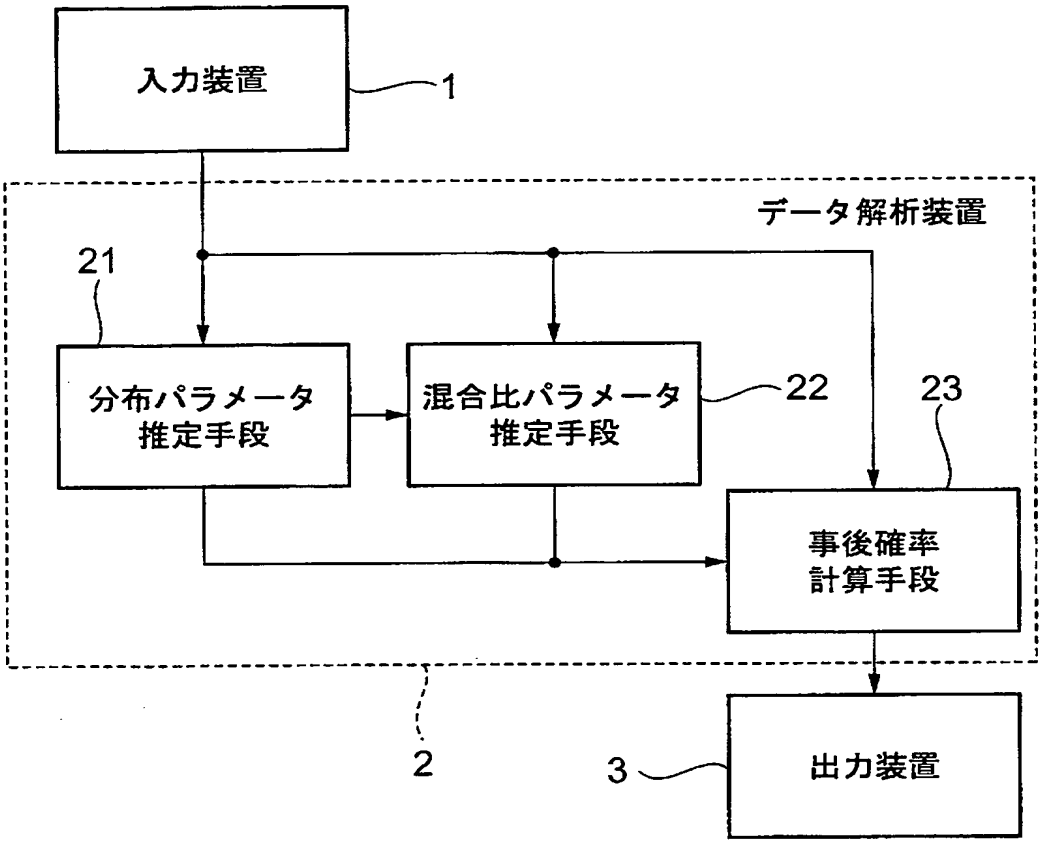
【符号の説明】

【００７１】
１　入力装置
２　データ解析装置
３　出力装置
４　記録媒体
５　データ解析装置
２１　分布パラメータ推定手段
２２　混合比パラメータ推定手段
２３　事後確率計算手段

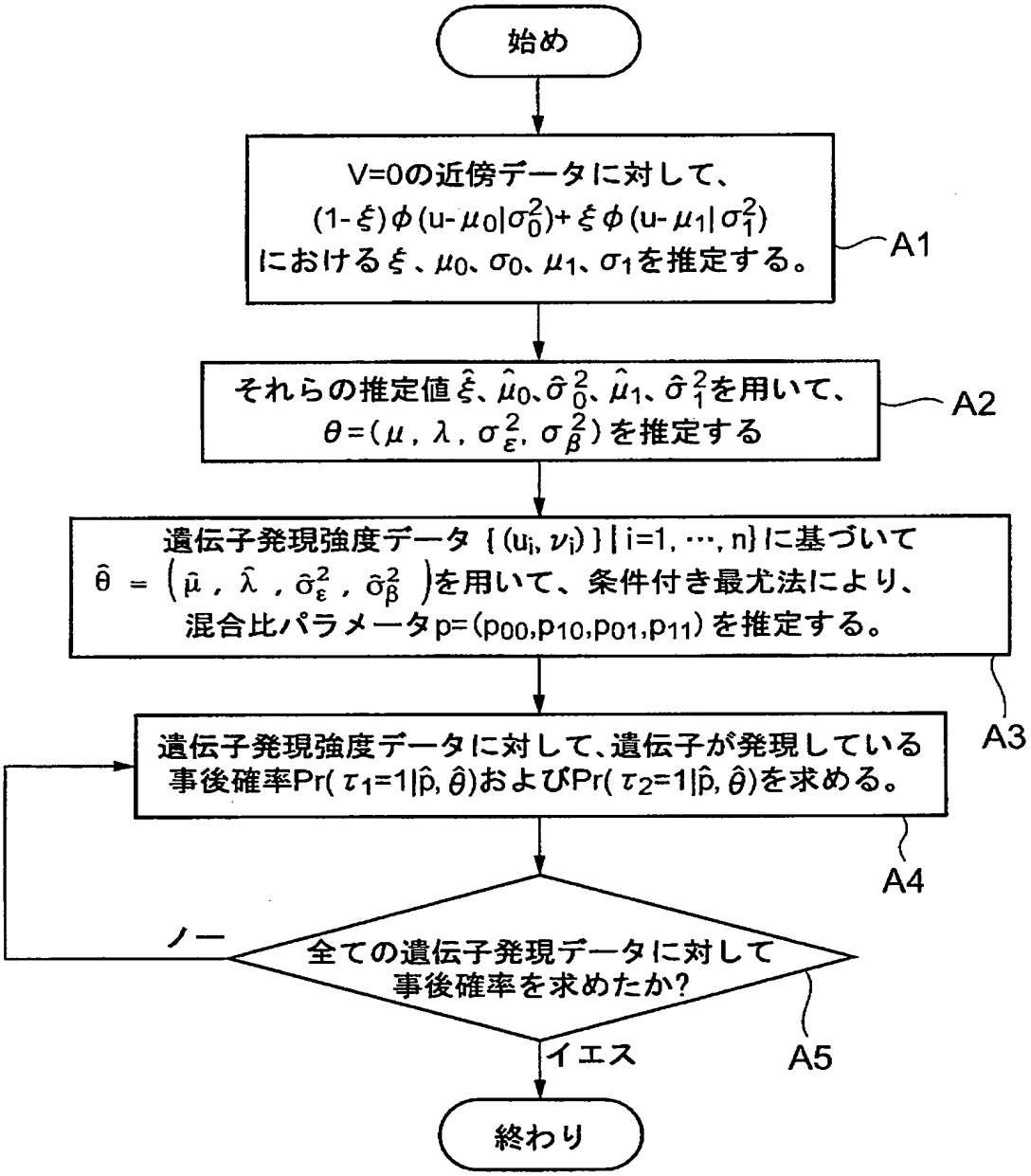
【書類名】図面
【図１】



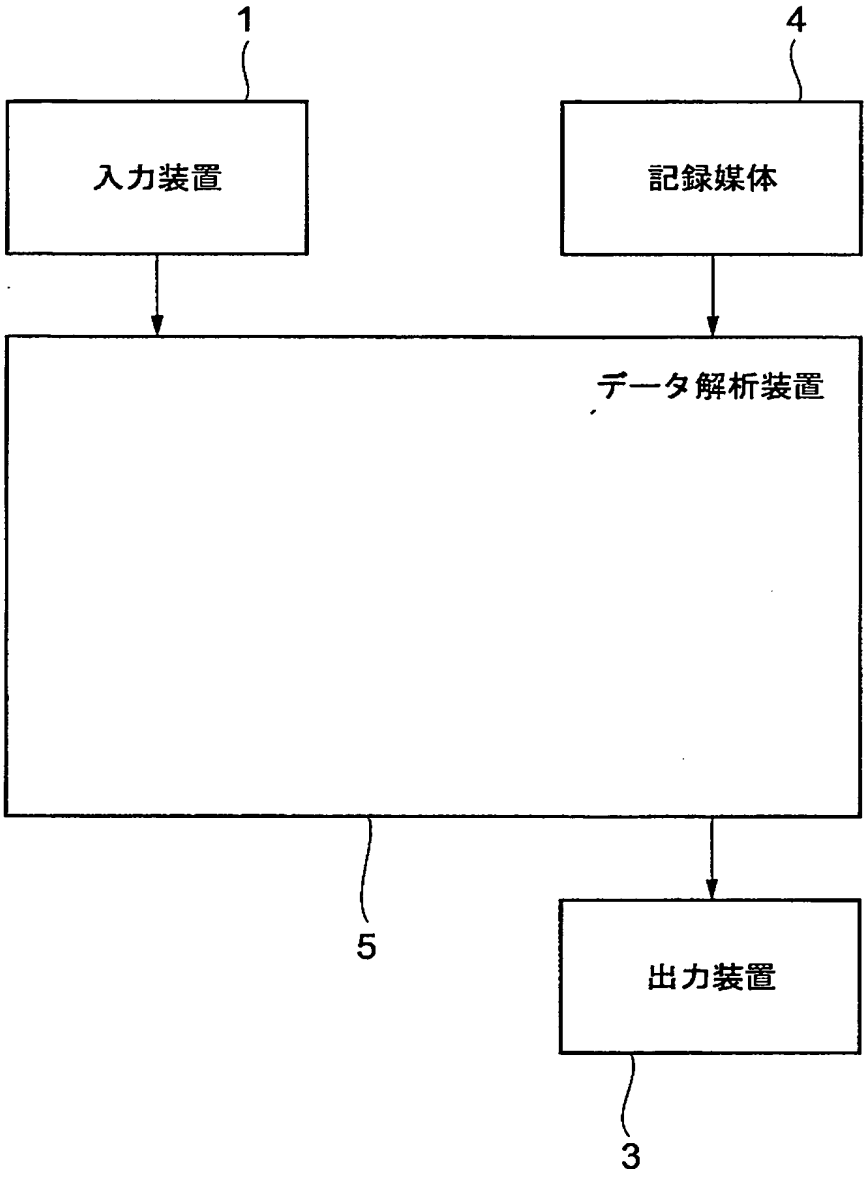
【図２】


入力装置
1
データ解析装置
21
分布パラメータ
推定手段
混合比パラメータ
推定手段
22
23
事後確率
計算手段
2
3
出力装置

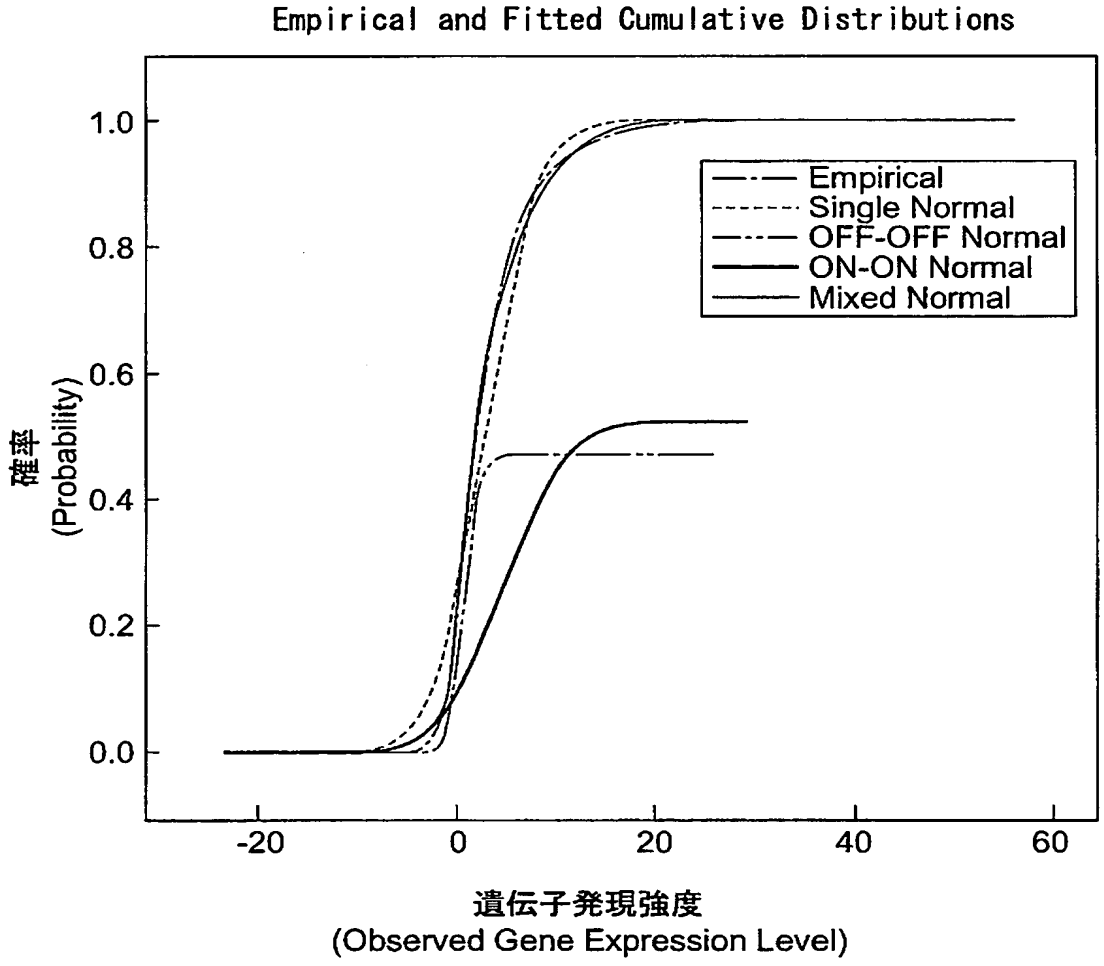
【図３】


始め
V=0の近傍データに対して、
$(1-\xi)\phi(u-\mu_0|\sigma_0^2)+\xi\phi(u-\mu_1|\sigma_1^2)$
におけるξ、μ_0、σ_0、μ_1、σ_1を推定する。
A1
それらの推定値$\hat{\xi}$、$\hat{\mu}_0$、$\hat{\sigma}_0^2$、$\hat{\mu}_1$、$\hat{\sigma}_1^2$を用いて、
$\theta=(\mu, \lambda, \sigma_\varepsilon^2, \sigma_\beta^2)$を推定する
A2
遺伝子発現強度データ $\{(u_i, v_i)\}\,|\,i=1, \cdots, n\}$ に基づいて
$\hat{\theta} = (\hat{\mu}, \hat{\lambda}, \hat{\sigma}_\varepsilon^2, \hat{\sigma}_\beta^2)$を用いて、条件付き最尤法により、
混合比パラメータ$p=(p_{00},p_{10},p_{01},p_{11})$を推定する。
A3
遺伝子発現強度データに対して、遺伝子が発現している
事後確率$Pr(\tau_1=1|\hat{p}, \hat{\theta})$および$Pr(\tau_2=1|\hat{p}, \hat{\theta})$を求める。
A4
ノー
全ての遺伝子発現データに対して
事後確率を求めたか?
A5
イエス
終わり

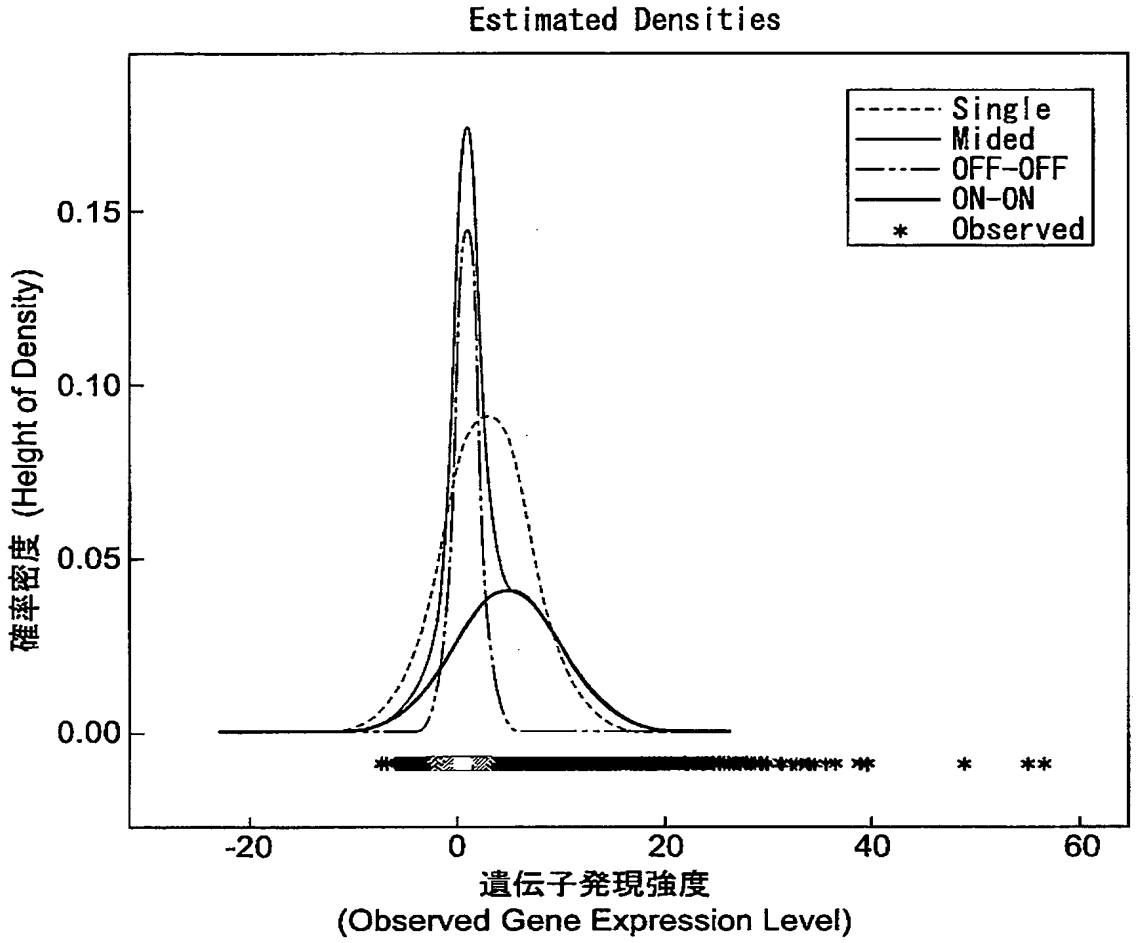
【図４】



【図５】


Empirical and Fitted Cumulative Distributions
1.0
0.8
0.6
0.4
0.2
0.0
確率
(Probability)
-20
0
20
40
60
遺伝子発現強度
(Observed Gene Expression Level)
Empirical
Single Normal
OFF-OFF Normal
ON-ON Normal
Mixed Normal

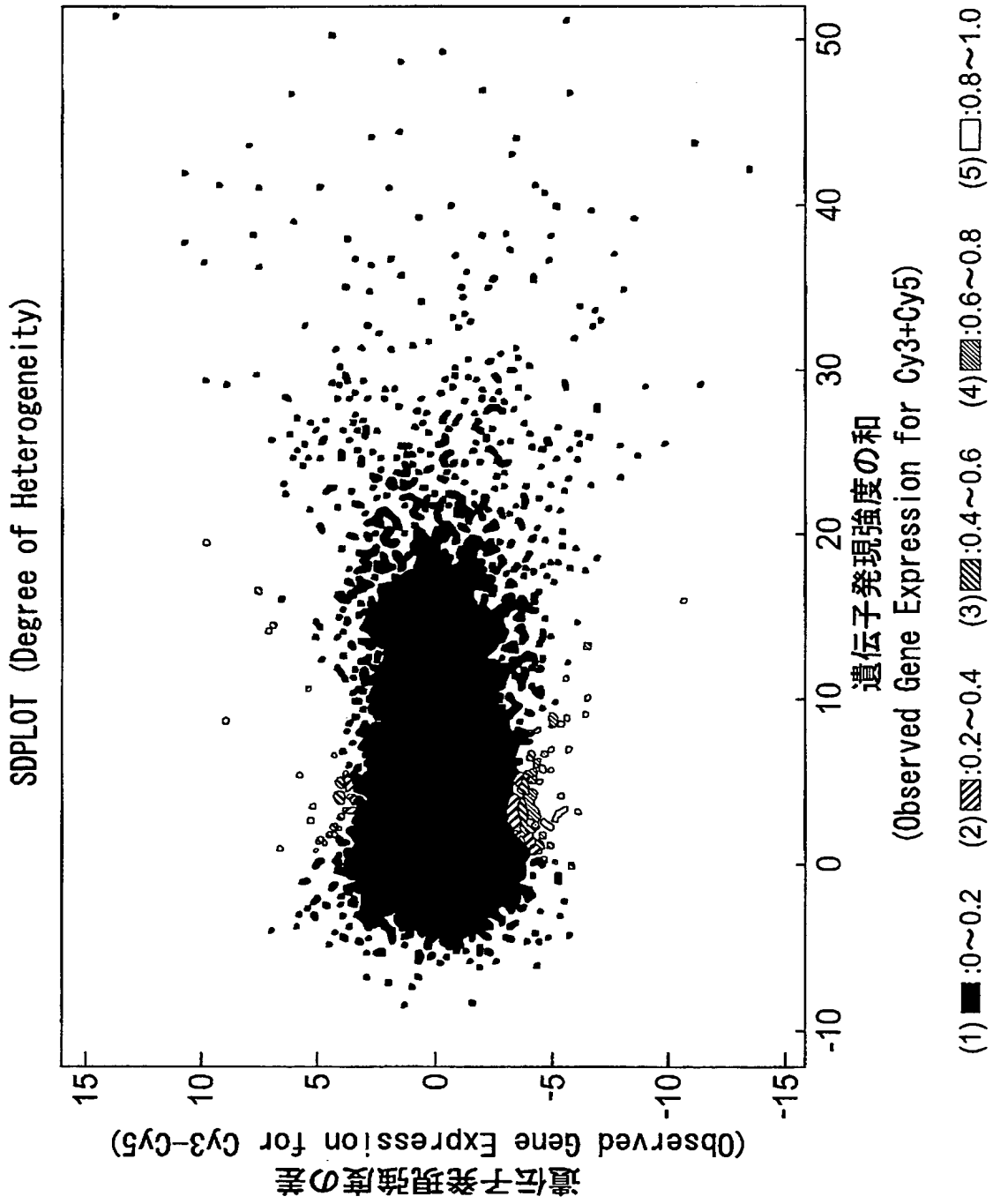
【図６】



(1) ■:0～0.2　(2) ▧:0.2～0.4　(3) ▨:0.4～0.6

(4) ▨:0.6～0.8　(5) □:0.8～1.0

【図７】


SDPLOT (Degree of Heterogeneity)
遺伝子発現強度の差
(Observed Gene Expression for Cy3-Cy5)
15
10
5
0
-5
-10
-15
-10
0
10
20
30
40
50
遺伝子発現強度の和
(Observed Gene Expression for Cy3+Cy5)
(1)■:0～0.2
(2)▨:0.2～0.4
(3)▨:0.4～0.6
(4)▨:0.6～0.8
(5)□:0.8～1.0

【書類名】要約書
【要約】
【課題】マイクロアレイデータを用いた解析の精度および効率を高めるために、遺伝子の発現状況に関する真の信号と測定誤差の分離を行い、チャンネル毎に遺伝子が発現している確率を推定する。
【解決手段】本発明の遺伝子発現状況推定システムは、マイクロアレイデータを入力する入力装置１と、プログラム制御により動作するデータ解析装置２と、出力装置３とを含む。データ解析装置２は、入力装置１から与えられた遺伝子発現強度データを用いて、混合正規分布のコンポーネント毎の分布パラメータおよび混合比パラメータを推定するパラメータ推定手段２１，２２と、推定された各パラメータを用いて、チャンネル毎に遺伝子の発現状況に関する事後確率を計算する事後確率計算手段２３を有し、計算された事後確率は出力装置３に出力する。
【選択図】図２

特願２００３－２７５９８３

出　願　人　履　歴　情　報

識別番号　　［０００００４２３７］

１．変更年月日　　１９９０年　８月２９日
［変更理由］　　新規登録
住　所　　東京都港区芝五丁目７番１号
氏　名　　日本電気株式会社

特願２００３－２７５９８３

出　願　人　履　歴　情　報

識別番号　　［５０３０７７１６５］

１．変更年月日　　２００３年　２月２６日
［変更理由］　　新規登録
住　所　　広島県廿日市市宮園９丁目１の７
氏　名　　大瀧　慈

特願２００３－２７５９８３

出　願　人　履　歴　情　報

識別番号　　［５００５３５３０１］

１．変更年月日　　２０００年１１月２０日
［変更理由］　　新規登録
住　所　　東京都中央区八丁堀二丁目２６番９号　グランデビルディング
氏　名　　社団法人バイオ産業情報化コンソーシアム